Sequential Information Guided Sensing

Ruiyang Song, Yao Xie, Sebastian Pokutta

***Abstract*—We study the value of information in sequential compressed sensing by characterizing the performance of sequential information guided sensing in practical scenarios when information is inaccurate. In particular, we assume the signal distribution is parameterized through Gaussian or Gaussian mixtures with estimated mean and covariance matrices, and we can measure compressively through a noisy linear projection or using one-sparse vectors, i.e., observing one entry of the signal each time. We establish a set of performance bounds for the bias and variance of the signal estimator via posterior mean, by capturing the conditional entropy (which is also related to the size of the uncertainty), and the additional power required due to inaccurate information to reach a desired precision. Based on this, we further study how to estimate covariance based on direct samples or covariance sketching. Numerical examples also demonstrate the superior performance of Info-Greedy Sensing algorithms compared with their random and non-adaptive counterparts.**

***Index Terms*—compressed sensing, mutual information, sequential methods, sketching**

I. Introduction

Sequential compressed sensing is a promising new information acquisition and recovery technique to process big data that arises in various applications such as compressive imaging [1]–[3], power network monitoring [4], and large scale sensor networks [5]. The sequential nature of the problems is either because the measurements are taken one after another, or due to the fact that the data is obtained in a streaming fashion so that it has to be processed in one pass.

To harvest the benefits of adaptivity in sequential compressed sensing, various algorithms have been developed (see [6] for a review.) We may classify these algorithms as (1) being agnostic about the signal distribution and, hence, using random measurements [7]–[13]; (2) exploiting additional structure of the signal (such as graphical structure [14] and tree-sparse structure [15], [16]) to design measurements; (3) exploiting the distributional information of the signal in choosing the measurements possibly through maximizing mutual information: the seminal Bayesian compressive sensing work [17], Gaussian mixture models (GMM) [18], [19], and our earlier work [6] which presents a general framework for information guided sensing referred to as *Info-Greedy Sensing*.

Such additional knowledge about signal structure or distributions are various forms of *information* about the unknown signal. Information may play a distinguishing role: as the compressive imaging example demonstrated in Fig. 1 (see Section IV for more details), with a bit of (albeit inaccurate) information estimated via random samples of small patches of the image, Info-Greedy Sensing is able to recover details of a high-resolution image, whereas random measurements completely miss the image.

In this paper we examine the value of information in sequential compressed sensing by considering Info-Greedy Sensing when information is imprecise. Info-Greedy Sensing is a framework introduced in [6] that aims at designing subsequent measurements to maximize the mutual information conditioned on previous measurements. Conditional mutual information is a natural metric here, as it captures exclusively useful new information between the signal and the results of the measurements disregarding noise and what has already been learned from previous measurements. We assume information is parameterized imperfectly and captured through sample estimates or "sketching", and when measurements of the unknown signal are compressive or even one-sparse (we are only able to inspect one entry of the signal). As shown in [6], Info-Greedy Sensing for a Gaussian signal becomes a simple iterative algorithm: choosing the measurement as the leading eigenvector of the conditional signal covariance matrix in that iteration, and then update the covariance matrix via a simple rank-one update, or, equivalently, choosing measurement vectors $a_1, a_2, \ldots$ as the orthonormal eigenvectors of the signal covariance matrix Σ in a decreasing order of eigenvalues. This can also be easily generalized to GMM signals, where a heuristic that works well is to measure in the dominant eigenvector direction of the Gaussian component with the highest posterior weight in that iteration.

In practice, we may be able to estimate the signal covariance matrix to initialize the algorithm through a training session. For Gaussian signals, there are two possible approaches: either using training samples that are sampled from the same distribution, or through the so-called "covariance sketching" [20]–[22] based on low-dimensional random sketches of the samples. As a consequence, the measurement vectors are calculated from eigenvectors of the estimated covariance matrix $\widehat{\Sigma}$, which deviates from the optimal directions. Since we almost always have to use an estimate for the signal covariance, it is crucial to quantify the performance of sensing algorithms with model mismatch and shed some light on how to properly initialize the algorithm.

Ruiyang Song (Email: songry12@mails.tsinghua.edu.cn) is with the Dept. of Electronic Engineering, Tsinghua University. Yao Xie (Email: yao.xie@isye.gatech.edu) and Sebastian Pokutta (Email: sebastian.pokutta@isye.gatech.edu) are with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

This work is partially supported by an NSF CAREER Award CMMI-1452463 and an NSF grant CCF-1442635. Ruiyang Song was visiting the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology while working on this paper.

In this paper we characterize the performance of Info-Greedy Sensing for Gaussian and GMM signals (with possibly low-rank covariance matrices) when the true signal covariance matrix is replaced with a proxy, which may be an estimate from direct samples or using a covariance sketching scheme. We establish a set of theoretical results including (1) studying the bias and variance of the signal estimator via posterior mean, by relating the error in the covariance matrix $\|\Sigma - \widehat{\Sigma}\|$ to the entropy of the signal posterior distribution after each sequential measurement, (2) establishing an upper bound on the additional power needed to achieve the signal precision $\|x - \hat{x}\| \leq \varepsilon$; and (3) translate these into requirements on the choice of the sample covariance matrix through direct estimation or through covariance sketching. Furthermore, we also study Info-Greedy Sensing in a special setting when the measurement vector is desired to be one-sparse, and establish analogously a set of theoretical results. Such a requirement arises from applications such as nondestructive testing (NDT) [23] or network tomography. We also present numerical examples to demonstrate the superior performance of Info-Greedy Sensing compared to a batch method (where measurements are not adaptive) when there is mismatch.

Our notations are standard. Denote $[n] \triangleq \{1, 2, \ldots, n\}$; $\|X\|$, $\|X\|_F$, and $\|X\|_*$ represent the spectral norm, the Frobenius norm, and the nuclear norm of a matrix X, respectively; let $\nu_i(\Sigma)$ denote the ith largest eigenvalue of a positive semi-definite matrix Σ; $\|x\|_0$, $\|x\|_1$, and $\|x\|$ represent the ℓ_0, ℓ_1 and ℓ_2 norm of a vector x, respectively; let χ^2_n be the quantile function of the chi-squared distribution with n degrees of freedom; let $\mathbb{E}[x]$ and $\mathrm{Var}[x]$ denote the mean and the variance of a random variable x; we write $X \succeq 0$ to indicate that the matrix is positive semi-definite; $\phi(x|\mu, \Sigma)$ denotes the probability density function of the multi-variate Gaussian with mean μ and covariance matrix Σ; let e_j denote the jth column of identity matrix I (i.e., e_j is a vector with only one non-zero entry at location j); and $(x)^+ = \max\{x, 0\}$ for $x \in \mathbb{R}$.

II. Info-Greedy Sensing

A typical sequential compressed sensing setup is as follows. Let $x \in \mathbb{R}^n$ be an unknown n-dimensional signal. We make K measurements of x sequentially

$$y_k = a_k^\mathsf{T} x + w_k, \quad k = 1, \ldots, K,$$

and the power of the measurement is $\|a_k\|^2 = \beta_k$. The goal is to recover x using measurements $\{y_k\}_{k=1}^K$. Consider a Gaussian signal $x \sim \mathcal{N}(0, \Sigma)$ with known zero mean and covariance matrix Σ (here without loss of generality we have assumed the signal has zero mean). Assume the rank of Σ is s and the signal can be low-rank $s \ll n$ (however, the algorithm does not require the covariance to be low-rank). Info-Greedy Sensing [6] chooses each measurement to maximize the conditional mutual information

$$a_k \leftarrow \operatorname*{argmax}_a \mathbb{I}\left[x; a^\mathsf{T} x + w \,|\, y_j, a_j, j < k\right] / \beta_k. \qquad (1)$$

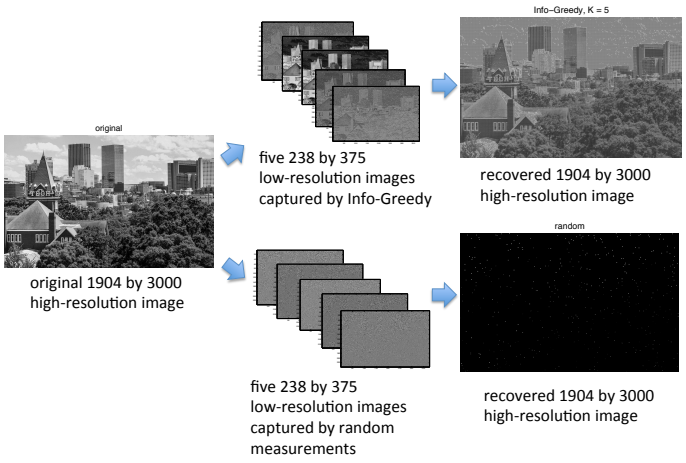


Fig. 1: Value of information in sensing a high-resolution image of size 1904×3000. Here, compressive linear measurements correspond to extracting the so-called *features* in compressive imaging [1]–[3]. In this example, the compressive imaging system captures 5 low resolution images of size 238-by-275 using masks designed by Info-Greedy Sensing or random masks (this corresponds to compressing the data into 8.32% of its original dimensionality). Info-Greedy Sensing performs much better than random features and preserves richer details in the recovered image. Details are explained in Section IV-C2.

The goal is to use a minimum number of measurements (or total power) so that the estimated signal is recovered with precision ε; i.e., $\|\widehat{x} - x\| < \varepsilon$ with a high probability p. Define

$$\chi_{n,p,\varepsilon} \triangleq \varepsilon^2 / \chi^2_n(p),$$

and we will show in the following that this is a fundamental quantity that determines the termination condition of our algorithm to achieve the precision ε with the confidence level p. Note that $\chi_{n,p,\varepsilon}$ is a precision ε adjusted by the confidence level.

A. Gaussian signal

In [6], we have devised a solution to (1) when the signal is Gaussian. The measurement will be made in the directions of the eigenvectors of Σ in a decreasing order of eigenvalues, and the powers (or the number of measurements) will be such that the eigenvalues after the measurements are sufficiently small (i.e., less than ε). The power allocation depends on the noise variance, signal recovery precision ε, and confidence level p, as given in Algorithm 1.

B. Gaussian mixture model (GMM) signals

The probability density function of GMM is given by

$$p(x) = \sum_{c=1}^{C} \pi_c \phi(x|\mu_c, \Sigma_c),$$

where C is the number of classes, and π_c is the probability that sample is drawn from class c. Unlike for Gaussian signals, the

Algorithm 1 Info-Greedy Sensing for Gaussian signals

Require: assumed signal mean μ and covariance matrix Σ, noise variance σ^2, recovery accuracy ε, confidence level p

repeat
  $(\lambda, u) \leftarrow$ largest eigenvalue and associated normalized eigenvector of Σ
  $\beta \leftarrow \sigma^2(1/\chi_{n,p,\varepsilon} - 1/\lambda)^+$
  $a = \sqrt{\beta}u$, $y = a^\intercal x + w$ {measure}
  $\mu \leftarrow \mu + \Sigma a(y - a^\intercal \mu)/(\beta\lambda + \sigma^2)$ {mean}
  $\Sigma \leftarrow \Sigma - \Sigma a a^\intercal \Sigma/(\beta\lambda + \sigma^2)$ {covariance}
until $\|\Sigma\| \leq \chi_{n,p,\varepsilon}$ {all eigenvalues small}
return signal estimate $\hat{x} = \mu$

mutual information of GMM has no explicit form. However, for GMM signals, there are two approaches that tend to work well: Info-Greedy Sensing derived based on a gradient descent approach [6], [19] uses the fact that the gradient of the conditional mutual information with respect to a is a linear transform of the minimum mean square error (MMSE) matrix [24], [25], and the so-called *greedy heuristic* which approximately maximizes the mutual information. The greedy heuristic picks the Gaussian component with the highest posterior π_c at that moment, and chooses the next measurement a to be its eigenvector associated with the maximum eigenvalue, as summarized in Algorithm 2. The greedy heuristic can be implemented more efficiently compared to the gradient descent approach and sometimes have competitive performance (see, e.g. [6]).

Algorithm 2 Heuristic Info-Greedy Sensing for GMM signals

Require: number of components C, assumed means $\{\mu_c\}$, covariances $\{\Sigma_c\}$, initial weights $\{\pi_c\}$, noise variance σ^2, confidence level p

repeat
  $c^* \leftarrow \arg\max_c \pi_c$
  $(\lambda, u) \leftarrow$ largest eigenvalue and associated normalized eigenvector of Σ_{c^*}
  $\beta \leftarrow \sigma^2(1/\chi_{n,p,\varepsilon} - 1/\lambda)^+$
  $a = \sqrt{\beta}u$, $y = a^\intercal x + w$ {measure}
  for $c = 1, \ldots, C$ **do**
    $\mu_c \leftarrow \mu_c + [(y - a^\intercal \mu_c)/(a^\intercal \Sigma_c a + \sigma^2)]\Sigma_c a$
    $\Sigma_c \leftarrow \Sigma_c - \Sigma_c a a^\intercal \Sigma_c/(a^\intercal \Sigma_c a + \sigma^2)$
    $\pi_c \leftarrow K\pi_c \exp\{-\frac{1}{2}(y - a^\intercal \mu_c)^2/(a^\intercal \Sigma_c a + \sigma^2)\}$
    (K: normalizing constant)
  end for
until $\|\Sigma_{c^*}\| \leq \chi_{n,p,\varepsilon}$
return signal class $c^* = \arg\max_c \pi_c$, estimate $\hat{x} = \mu_{c^*}$

C. One-sparse measurement

The problem of Info-Greedy Sensing with sparse measurement constraint, i.e., each measurement has only k_0 non-zero entries $\|a\|_0 = k_0$, has been examined in [6] and solved using outer approximation (cutting planes). Here we will focus on one-sparse measurements, $\|a\|_0 = 1$, as it is an important instance arising in applications such as nondestructive testing (NDT).

Algorithm 3 Info-Greedy Sensing with sparse measurement $\|a\|_0 = 1$, for Gaussian signals

Require: assumed signal mean μ and covariance matrix Σ, noise variance σ^2, recovery accuracy ε, confidence level p

repeat
  $j^* \leftarrow \arg\max_j \Sigma_{jj}$
  $a \leftarrow \sqrt{\beta}e_{j^*}$, $y = a^\intercal x + w$ {measure}
  $\mu \leftarrow \mu + \Sigma a(y - a^\intercal \mu)/(\beta\Sigma_{j^*j^*} + \sigma^2)$ {mean}
  $\Sigma \leftarrow \Sigma - \Sigma a a^\intercal \Sigma/(\beta\Sigma_{j^*j^*} + \sigma^2)$ {covariance}
until $\|\Sigma\| \leq \chi_{n,p,\varepsilon}$ {all eigenvalues small}
return signal estimate $\hat{x} = \mu$

Info-Greedy Sensing with one-sparse measurements can be readily derived. Note that the mutual information between x and the outcome using one-sparse measurement $y_1 = e_j^\intercal x + w_1$ is given by

$$\mathbb{I}[x; y_1] = \frac{1}{2}\ln(\Sigma_{jj}/\sigma^2 + 1),$$

where Σ_{jj} denote the jth diagonal entry of matrix Σ. Hence, the measurement that maximizes the mutual information is given by e_{j^*} where $j^* = \arg\max_j \Sigma_{jj}$, i.e., measuring in the signal coordinate with the largest variance or largest uncertainty. Then Info-Greedy Sensing measurements can be found iteratively, as presented in Algorithm 3. Note that the correlation of signal coordinates are reflected in the update of the covariance matrix: if the ith and jth coordinates of the signal are highly correlated, then the uncertainty in j will also be greatly reduced if we measure in i. A similar algorithm with one-sparse measurement for GMM signals can be derived, where in each iteration we select the component with the largest weight and measure in the signal coordinate with largest variance.

D. Updating covariance with sequential data

If our goal is to estimate a sequence of data $x_1, x_2, \ldots$ (versus just estimating a single instance), we may be able to update the covariance matrix using the already estimated signals simply via

$$\widehat{\Sigma}_t = \alpha\widehat{\Sigma}_{t-1} + (1 - \alpha)\hat{x}_t\hat{x}_t^\intercal, \quad t = 1, 2, \ldots, \tag{2}$$

and the initial covariance matrix is specified by our prior knowledge $\widehat{\Sigma}_0 = \widehat{\Sigma}$. Using the updated covariance matrix $\widehat{\Sigma}_t$, we design the next measurement for signal x_{t+1}. This way we may be able to correct the inaccuracy of $\widehat{\Sigma}$ by including new samples. We refer to this method as "Info-Greedy-2" hereafter.

III. Performance bounds

In the following, we establish performance bounds, for cases when we (1) sense Gaussian and GMM signals using estimated covariance matrices; (2) sense Gaussian signals with one-sparse measurements.

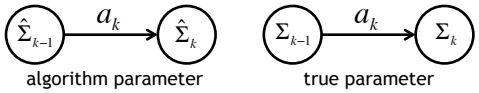


Fig. 2: Parameter updates performed by the algorithm and updates happen on the true distribution.

A. Gaussian case with model mismatch

To analyze the performance of our algorithms when the assumed covariance $\widehat{\Sigma}$ used in Algorithm 1 is different from the true signal covariance matrix Σ, we introduce the following notations. Let the eigenpairs of Σ with the eigenvalues (which can be zero) ranked from the largest to the smallest to be $(\lambda_1, u_1), (\lambda_2, u_2), \ldots, (\lambda_n, u_n)$, and let the eigenpairs of $\widehat{\Sigma}$ with the eigenvalues (which can be zero) ranked from the largest to the smallest to be $(\hat{\lambda}_1, \hat{u}_1), (\hat{\lambda}_2, \hat{u}_2), \ldots, (\hat{\lambda}_n, \hat{u}_n)$. Let the updated covariance matrix in Algorithm 1 starting from $\widehat{\Sigma}$ after k measurements be $\widehat{\Sigma}_k$, and the true posterior covariance matrix of the signal conditioned on these measurements be Σ_k. The relations of these notations are illustrated in Fig. 2.

Note that since each time we measure in the direction of the dominating eigenvector of the posterior covariance matrix, $(\hat{\lambda}_k, \hat{u}_k)$ and (λ_k, u_k) correspond to the largest eigenpair of $\widehat{\Sigma}_{k-1}$ and Σ_{k-1}, respectively. Furthermore, define the difference between the true and the assumed conditional covariance matrices after k measurements as

$$E_k = \widehat{\Sigma}_k - \Sigma_k, \quad k = 1, \ldots, K,$$

and their sizes

$$\delta_k = \|E_k\|, \quad k = 1, \ldots, K.$$

Let the eigenvalues of E_k be $e_1 \geq e_2 \geq \cdots \geq e_n$; then $\delta_k = \max\{|e_1|, |e_n|\}$. Let

$$\delta_0 = \|\widehat{\Sigma} - \Sigma\|$$

denote the size of the initial mismatch.

1) Deterministic mismatch: First we assume the mismatch is deterministic, and find bounds for bias and variance of the estimated signal. Assume the initial mean is $\hat{\mu}$ and the true signal mean is μ, the updated mean using Algorithm 1 after k measurements is $\hat{\mu}_k$, and the true posterior mean is μ_k.

Theorem 1 (Unbiasedness)**.** *After k measurements, the expected difference between the updated mean and the true posterior mean is given by*

$$\mathbb{E}[\hat{\mu}_k - \mu_k] = (\hat{\mu} - \mu) \cdot \prod_{j=1}^{k} (I_n - \frac{\beta_j \hat{\lambda}_j}{\beta_j \hat{\lambda}_j + \sigma^2} \hat{u}_j \hat{u}_j^\intercal).$$

Moreover, if $\hat{\mu} = \mu$, i.e., the assumed mean is accurate, the estimator is unbiased throughout all the iterations $\mathbb{E}[\hat{\mu}_k - \mu_k] = 0$, for $k = 1, \ldots, K$.

Next we show that the variance of the estimator, when the initial mismatch $\|\widehat{\Sigma} - \Sigma\|$ is sufficiently small, reduces gracefully. This is captured through the reduction of entropy, which is also a measure of the uncertainty in the estimator. In particular, we consider the posterior entropy of the signal conditioned on the previous measurement outcomes. Since the entropy of a Gaussian signal $x \sim \mathcal{N}(\mu, \Sigma)$ is given by $\mathbb{H}[x] = \ln\left[(2\pi e)^{n/2} \det^{1/2}(\Sigma)\right]$, the conditional mutual information is the log of the determinant of the conditional covariance matrix, or equivalently the log of the volume of the ellipsoid defined by the covariance matrix. Here, to accommodate the scenario where the covariance matrix is low-rank (our earlier assumption), we consider a modified definition for conditional entropy, which is the log of the volume of the ellipsoid on the low-dimensional space that the signal lies on:

$$\mathbb{H}[x \,|\, y_j, a_j, j \leq k] = \ln[(2\pi e)^{s/2} \mathsf{Vol}(\Sigma_k)],$$

where $\mathsf{Vol}(\Sigma_k)$ is the volume of the ellipse defined by Σ_k equal to the product of its non-zero eigenvalues.

Theorem 2 (Entropy of estimator)**.** *If for some constant $\delta \in (0, 1)$ the error satisfies*

$$\|\widehat{\Sigma} - \Sigma\| \leq \frac{\delta}{4^{K+1}} \chi_{n,p,\varepsilon},$$

then for $k = 1, \ldots, K$,

$$\mathbb{H}[x \,|\, y_j, a_j, j \leq k] \leq \frac{s}{2} \left\{ \ln[2\pi e \ \mathrm{tr}(\Sigma)] - \sum_{j=1}^{k} \ln(1/f_j) \right\}, \tag{3}$$

where

$$f_k = 1 - \frac{1 - \delta}{s} \frac{\beta_k \hat{\lambda}_k}{\beta_k \hat{\lambda}_k + \sigma^2} \in (0, 1), \quad k = 1, \ldots, K. \tag{4}$$

In the proof of Theorem 2, we use the trace of the underlying actual covariance matrix $\mathrm{tr}(\Sigma_k)$ as potential function, which serves as a surrogate for the product of eigenvalues that determines the volume of the ellipsoid and hence the entropy, since it is much easier to calculate the trace of the observed covariance matrix $\mathrm{tr}(\widehat{\Sigma}_k)$. The following recursion is crucial for the derivation: for an assumed covariance matrix Σ, after measuring in the direction of a unit norm eigenvector u with eigenvalue λ using power β, the updated matrix takes the form of

$$\begin{aligned} &\Sigma - \Sigma\sqrt{\beta}u \left(\sqrt{\beta}u^\intercal \Sigma \sqrt{\beta}u + \sigma^2\right)^{-1} \sqrt{\beta}u^\intercal \Sigma \\ &= \frac{\lambda\sigma^2}{\beta\lambda + \sigma^2} uu^\intercal + \Sigma^{\perp u}, \end{aligned} \tag{5}$$

where $\Sigma^{\perp u}$ is the component of Σ in the orthogonal complement of u. Thus, the only change in the eigen-decomposition of Σ is the update of the eigenvalue of u from λ to $\lambda\sigma^2/(\beta\lambda + \sigma^2)$. Based on (5), after one measurement, the trace of the covariance matrix becomes

$$\mathrm{tr}(\widehat{\Sigma}_k) = \mathrm{tr}(\widehat{\Sigma}_{k-1}) - \frac{\beta_k \hat{\lambda}_k^2}{\beta_k \hat{\lambda}_k + \sigma^2}. \tag{6}$$

Remark 1. *The upper bound of the posterior signal entropy in (3) shows that the amount of uncertainty reduction by the kth measurement is roughly $(s/2)\ln(1/f_k)$.*

Remark 2. *Use the inequality $\ln(1-x) \le -x$ for $x \in (0,1)$, we have that in (3)*

$$\mathbb{H}\left[x \,|\, y_j, a_j, j \le k\right] \le \frac{s}{2}\ln[2\pi e \mathrm{tr}(\Sigma)] - \frac{1-\delta}{2}\sum_{j=1}^{k}\frac{\beta_j\hat{\lambda}_j}{\beta_j\hat{\lambda}_j+\sigma^2}$$

$$= \frac{s}{2}\ln[2\pi e \mathrm{tr}(\Sigma)] - \frac{k(1-\delta)}{2} + \frac{(1-\delta)}{2}\sum_{j=1}^{k}\frac{\chi_{n,p,\varepsilon}}{\hat{\lambda}_j}.$$

On the other hand, in the ideal case if the true covariance matrix is used, the posterior entropy of the signal is given by

$$\mathbb{H}_{\text{ideal}}\left[x, |y_j, a_j, j \le k\right] = \frac{1}{2}\ln[(2\pi e)^s\prod_{j=1}^{s}\lambda_j] - \frac{1}{2}\sum_{j=1}^{k}\frac{\lambda_j}{\chi_{n,p,\varepsilon}}, \tag{7}$$

where $\tilde{\beta}_j = (1/\chi_{n,p,\varepsilon} - 1/\lambda_j)^+\sigma^2$. Hence, we have

$$\begin{aligned}\mathbb{H}\left[x \,|\, y_j, a_j, j \le k\right] &\le \mathbb{H}_{\text{ideal}}\left[x, |y_j, a_j, j \le k\right] + C \\ &\quad - \frac{1}{2}\sum_{j=1}^{k}\left[\frac{\lambda_j}{\chi_{n,p,\varepsilon}} + (1-\delta)\left(1 - \frac{\chi_{n,p,\varepsilon}}{\hat{\lambda}_j}\right)\right].\end{aligned} \tag{8}$$

where $C = \frac{s}{2}\ln[\mathrm{tr}(\Sigma)/\sqrt[s]{\prod_{j=1}^{s}\lambda_j}]$ is a constant independent of measurements. This upper bound has a nice interpretation: it characterizes the amount of uncertainty reduction with each measurement. For example, when the number of measurements required when using the assumed covariance matrix versus using the true covariance matrix are the same, we have $\lambda_j \ge \chi_{n,p,\varepsilon}$ and $\hat{\lambda}_j \ge \chi_{n,p,\varepsilon}$. Hence, the third term in (8) is upper bounded by $-k/2$, which means that the amount of reduction in entropy is roughly 1/2 nat per measurement.

Remark 3. *Consider the special case where the errors only occur in the eigenvalues of the matrix but not in the eigenspace U, i.e., $\widehat{\Sigma}-\Sigma = U diag\{e_1, \cdots, e_s\}U^\intercal$ and $\max_{1\le j\le s}|e_j| = \delta_0$, then the upper bound in (7) can be further simplified. Suppose only the first K $(K \le s)$ largest eigenvalues of $\widehat{\Sigma}$ are larger than the stopping criterion $\chi_{n,p,\varepsilon}$ required by the precision, i.e., the algorithm takes K iterations in total. Then*

$$\begin{aligned}\mathbb{H}\left[x \,|\, y_j, a_j, j \le k\right] \le \mathbb{H}_{\text{ideal}}&\left[x, |y_j, a_j, j \le k\right] \\ &+ K\ln(1+\delta_K/\chi_{n,p,\varepsilon}) \\ &+ \sum_{j=K+1}^{s}\ln(1+(\delta_0+\delta_K)/\lambda_j).\end{aligned}$$

The additional entropy relative to the ideal case $\mathbb{H}_{\text{ideal}}$ is typically small, because $\delta_K \le \delta_0 4^K$ (according to Lemma 7 in the appendix), δ_0 is on the order of ε^2, and hence the second term in the appendix is on the order of K^2; the third term will be small because δ_0 and δ_K are small compare to λ_j.

Note that, however, if the power allocations β_i are calculated using the eigenvalues of the assumed covariance matrix $\widehat{\Sigma}$, after $K = s$ iterations, we are not guaranteed to reach the desired precision ε with probability p. However, this becomes possible if we increase the total power slightly. The following theorem establishes an upper bound on the amount of extra total power needed to reach the same precision ε compared to the total power P_{ideal} if we use the correct covariance matrix.

Theorem 3 (Additional power required)**.** *Assume $K \le s$ eigenvalues of Σ are larger than $\chi_{n,p,\varepsilon}$. If*

$$\|\widehat{\Sigma} - \Sigma\| \le \frac{1}{4^{s+1}}\chi_{n,p,\varepsilon},$$

then to reach a precision ε at confidence level p, the total power P_{mismatch} required by Algorithm 1 when using $\widehat{\Sigma}$ is upper bounded by

$$P_{\text{mismatch}} < P_{\text{ideal}} + \left[\frac{20}{51}s + \frac{1}{272}K\right]\frac{\sigma^2}{\chi_{n,p,\varepsilon}}.$$

Remark 4. *In a special case when $K = s$ eigenvalues of Σ are larger than $\chi_{n,p,\varepsilon}$, under the conditions of Theorem 3, we have a simpler expression for the upper bound*

$$P_{\text{mismatch}} < P_{\text{ideal}} + \frac{323}{816}\frac{\sigma^2}{\chi_{n,p,\varepsilon}}s.$$

Note that the additional power required is quite small and is only linear in s. All other parameters are independent of the input matrix.

2) Initialize $\widehat{\Sigma}$: In the following we present schemes to estimate $\widehat{\Sigma}$ to reach the desired precision in Theorem 2: (1) using sample covariance matrix if we are able to obtain full dimensional training samples; (2) using covariance sketching to estimate the covariance using random projections of the full dimensional training samples.

Suppose the sample covariance matrix is obtained from training samples $\tilde{x}_1, \ldots, \tilde{x}_L$ that are drawn i.i.d. from $\mathcal{N}(0,\Sigma)$, and $\widehat{\Sigma} = (1/L)\sum_{\ell=1}^{L}\tilde{x}_\ell\tilde{x}_\ell^\intercal$. Then we need L to be sufficiently large to reach the desired precision. The following Lemma 1 arises from a simple tail probability bound of the Wishart distribution (since the sample covariance matrix follows a Wishart distribution).

Lemma 1 (Initialize with sample covariance matrix)**.** *For any constant $\delta > 0$, we have $\|\widehat{\Sigma}-\Sigma\| \le \delta$ with probability exceeding $1 - 2n\exp(-\sqrt{n})$, as long as*

$$L \ge 4n^{1/2}\mathrm{tr}(\Sigma)(\|\Sigma\|/\delta^2 + 4/\delta).$$

Lemma 1 shows that the number of measurements needed to reach a precision δ for a sample covariance matrix is $\mathcal{O}(1/\delta^2)$ as expected.

We may also use a covariance sketching scheme similar to that described in [20]–[22] to estimate $\widehat{\Sigma}$. Covariance sketching is based on random projections of each training samples, and hence it is memory efficient when we are not able to store or operate

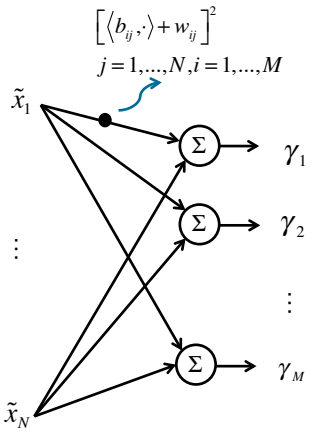


Fig. 3: Diagram of covariance sketching in our setting. The circle aggregates quadratic sketches from branches and computes the average.

on the full vectors directly. The covariance sketching scheme is described below and illustrated in Fig. 3. Assume training samples $\tilde{x}_i$, $i = 1, \ldots, N$ are drawn from the signal distribution. Each sample, $\tilde{x}_i$ is sketched M times using random sketching vectors b_{ij}, $j = 1, \ldots, M$, through a noisy linear measurement $(b_{ij}^{\mathsf{T}} x_i + w_{ijl})^2$, and we repeat this for L times ($l = 1, \ldots, L$) and compute the average energy to suppress noise[1]. This sketching process can be shown to be a linear operator $\mathcal{B}$ applied on the original covariance matrix Σ, as shown in Appendix A. We may recover the original covariance matrix from the vector of sketching outcomes $\gamma \in \mathbb{R}^M$ by solving the following convex optimization problem

$$\begin{aligned} \widehat{\Sigma} = \operatorname{argmin}_X \quad & \operatorname{tr}(X) \\ \text{subject to} \quad & X \succeq 0, \ \|\gamma - \mathcal{B}(X)\|_1 \le \tau, \end{aligned} \tag{9}$$

where τ is a user parameter that depends on the noise level. In the following theorem, we further establish conditions on the covariance sketching parameters N, M, L, and τ so that the recovered covariance matrix $\widehat{\Sigma}$ may reach the required precision in Theorem 2, by adapting the results in [22].

Lemma 2 (Initialize with covariance sketching)**.** *For any $\delta > 0$ the solution to (9) satisfies $\|\widehat{\Sigma} - \Sigma\| \le \delta$, with probability exceeding $1 - 2/n - 2/\sqrt{n} - 2n\exp(-\sqrt{n}) - \exp(-c_1 M)$, as long as the parameters M, N, L and τ satisfy the following conditions*

$$M > c_0 n s, \tag{10}$$

$$N \ge 4n^{1/2}\operatorname{tr}(\Sigma)\left(\frac{36M^2n^2\|\Sigma\|}{\tau^2} + \frac{24Mn}{\tau}\right), \tag{11}$$

$$L \ge \max\left\{\frac{M}{4n^2\|\Sigma\|}\sigma^2,\ \frac{1}{\sqrt{2[\operatorname{tr}(\Sigma)/\|\Sigma\|]Mn^2}}\sigma^2, \frac{6M}{\tau}\sigma^2\right\}, \tag{12}$$

$$\tau = M\delta/c_2, \tag{13}$$

where c_0, c_1, and c_2 are absolute constants.

B. Gaussian mixture model (GMM)

We also establish a lower bound on the number of measurements (or power) required to recover a GMM signal with high precision when there is model mismatch. The proof follows by identifying a connection between the Info-Greedy Sensing and the so-called *multiplicative weight update* (MWU) algorithms (see e.g., [26]–[28]). The MWU method is actually a meta-algorithm and its instantiations span a large family of algorithms. It has been re-derived under various names in various disciplines. MWU algorithms maintain a distribution over experts (which corresponds to different Gaussian components in our case) and form a solution by e.g., a majority vote or an average over the solutions suggested by the experts. (Here each Gaussian component will suggest a sensing vector.) The weights are updated in each round according to the posterior update. We will use the hedge version of MWU in deriving the result.

Theorem 4 (GMM with Mismatch)**.** *Denote the posterior mean and covariance of component c after k iterations as $\mu_{c,k}$ and $\Sigma_{c,k}$, and their perturbed counterparts as $\hat{\mu}_{c,k}$ and $\widehat{\Sigma}_{c,k}$, respectively. Let $\delta_{c,k} = \|\widehat{\Sigma}_{c,k} - \Sigma_{c,k}\|$, and m_c be the number of measurements (or power) required to ensure $\|x - \hat{x}\| < \varepsilon$ with probability p for a Gaussian signal $\mathcal{N}(\mu_c, \Sigma_c)$ corresponding to component c for all $c \in [C]$ if we start with sample covariance matrix $\widehat{\Sigma}_c$. Then if both the mismatch in the initial mean $|a^{\mathsf{T}}(\mu - \hat{\mu})|$ and the initial covariance $\|\widehat{\Sigma} - \Sigma\|$ are sufficiently small so that $\eta_0 = \mathcal{O}(\hat{\eta})$, then for a signal sampled from the c^*th component, we need at most*

$$\sum_{c=1}^{C} m_c + \mathcal{O}\left(\frac{\ln|C|}{\hat{\eta} + \eta_0}\right)$$

amount of power to ensure $\|x - \widehat{x}\| < \varepsilon$ with probability $p(1 - \hat{\eta} - 1/n)$ when sampling from the posterior distribution of π with probability. Here $\hat{\eta} = 1/2$,

$$\eta_0 = \frac{1}{\sigma^4} \cdot \max_{k=1}^{n}\left\{\sigma^2(|\varrho_k| + 2nb_k)|\varrho_k| + \beta_k n^2 b_k^2 \delta_{k-1}\right\},$$

$$\varrho_k \triangleq a_k^{\mathsf{T}}(\hat{\mu}_{c^*,k} - \mu_{c^*,k}),$$

$$b_k = \sqrt{(\lambda_{c^*,k} + \delta_{c^*,k-1}) + \sigma^2 + (a_k^{\mathsf{T}}(\mu_{c^*} - \mu_{c^*,k-1}))^2}.$$

[1]Our sketching scheme is slightly different from that used in [22] because we would like to use the square of the noisy linear measurements y_i^2 (where as the measurement scheme in [22] has a slightly different noise model). In practice, this means that we may use the same measurement scheme in the first stage as training to initialize the sample covariance matrix.

Note that here the constants are defined in terms of maximizing from 1 to n. This can be understood as if we run the algorithm until we have aquired n measurements.

Remark 5. *If our goal is to detect the correct component (rather than recovering the signal itself), we need at most* $\mathcal{O}\left(\frac{\ln|C|}{\hat{\eta}+\eta_0}\right)$ *samples if the true component is* c^*.

Remark 6. *Compared to GMM result without mismatch, which is on the order of* $\mathcal{O}(\ln|C|/\hat{\eta})$ *[6], this upper bound actually requires a smaller number of measurements* $\mathcal{O}(\ln|C|/(\hat{\eta}+\eta_0))$. *This is consistent with our intuition, and it says that if our estimation accuracy for the covariance matrices is low, then we should not "labor" as much. Because the estimation error will create an "error floor" which does not decrease by making more measurements, and it is not meaningful to make additional measurements below the noise floor. Of course, when there is covariance error, the overall error bound will be higher as well (which is already captured by a larger error bound).*

C. One-sparse measurement

In the following we provide performance bounds for the case of one-sparse measurements in Algorithm 3. Assume the signal covariance matrix is known precisely. Now that $\|a_k\|_0 = 1$, we have $a_k = \sqrt{\beta_k}u_k$, where $u_k \in \{e_1, \cdots, e_n\}$. Suppose the largest diagonal entry of $\Sigma^{(k-1)}$ is determined by

$$j_{k-1} = \arg\max_t \Sigma_{tt}^{(k-1)}.$$

From the update equation for the covariance matrix in Algorithm 3, the largest diagonal entry of $\Sigma^{(k)}$ can be determined from

$$j_k = \arg\max_t \left\{\Sigma_{tt}^{(k-1)} - \frac{(\Sigma_{tj_{k-1}}^{(k-1)})^2}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)} + \sigma^2/\beta_k}\right\}.$$

Let the correlation coefficient be denoted as

$$\rho_{ij}^{(k)} = \frac{(\Sigma_{ij}^{(k)})^2}{\Sigma_{ii}^{(k)}\Sigma_{jj}^{(k)}},$$

where the covariance of the ith and jth coordinate of x after k measurements is denoted as $\Sigma_{ij}^{(k)}$.

Lemma 3 (One sparse measurement. Recursion for trace of covariance matrix)**.** *Assume the minimum correlation for the* k*th iteration is* $\rho^{(k-1)} \in [0,1)$ *such that* $\rho^{(k-1)} \le |\rho_{ij_{k-1}}^{(k-1)}|$ *for any* $i \in [n]$. *Then for a constant* $\gamma > 0$, *if the power of the* k*th measurement* β_k *satisfies* $\beta_k \ge \sigma^2/\left(\gamma \max_t \Sigma_{tt}^{(k-1)}\right)$, *we have*

$$\mathrm{tr}(\Sigma_k) \le \left[1 - \frac{(n-1)\rho^{(k-1)}+1}{n(1+\gamma)}\right]\mathrm{tr}(\Sigma_{k-1}). \quad (14)$$

Lemma 3 provides a good bound for a one-step ahead prediction for the trace of the covariance matrix, as demonstrated in Fig. 4. Using the above lemma, we can obtain an upper bound on the number of measurements needed for one-sparse measurements.

Theorem 5 (Gaussian, one-sparse measurement)**.** *For constant* $\gamma > 0$, *when power is allocated satisfying* $\beta_k \ge \sigma^2/(\gamma \max_t \Sigma_{tt}^{(k-1)})$ *for* $k = 1, 2, \ldots, K$, *we have* $\|\hat{x} - x\| \le \varepsilon$ *with probability* p *as long as*

$$K \ge \max\left\{\frac{\ln[\mathrm{tr}(\Sigma)/\chi_{n,p,\varepsilon}]}{\ln\frac{1}{1-1/[n(1+\gamma)]}}, 0\right\}. \quad (15)$$

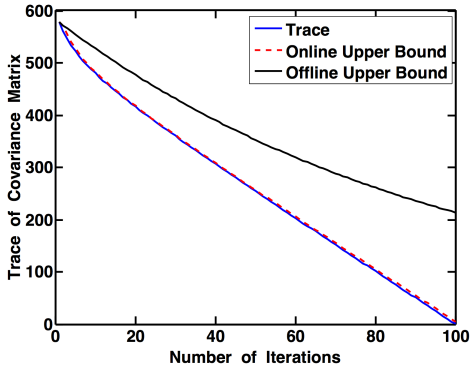


Fig. 4: One-step ahead prediction for the trace of the covariance matrix: the offline bound corresponds to applying (14) iteratively k times, and the online bound corresponds to predicting $\mathrm{tr}(\Sigma_k)$ using $\mathrm{tr}(\Sigma_{k-1})$. Here $n = 100$, $p = 0.95$, $\varepsilon = 0.1$, $\Sigma = dd^\mathsf{T} + 5I_n$ where $d = [1, \cdots, 1]^\mathsf{T}$.

The above theorem requires the number of iterations to be on the order of $\ln(1/\varepsilon)$ to reach precision ε (recall $\chi_{n,p,\varepsilon} = \varepsilon^2/\chi_n^2(p)$), as expected. It also suggest a method to allocate power: set β_k to be proportional to $\sigma^2/\max_t \Sigma_{tt}^{(k-1)}$: this captures the inter-dependence of the signal entries as these dependence will be affect the diagonal entries of the updated covariance matrix.

IV. Numerical examples

In the following, we have three sets of numerical examples to demonstrate the performance of Info-Greedy Sensing when there is mismatch in the signal covariance matrix, when the signal is sampled from Gaussian, and from GMM models, respectively.

A. Sensing Gaussian with mismatched covariance matrix

When the assumed covariance matrix for the signal x is equal to its true covariance matrix, Info-Greedy Sensing is identical to the batch method [19] (the batch method measures using the largest eigenvectors of the signal covariance matrix). However, when there is a mismatch between the two, Info-Greedy Sensing outperforms the batch method due to its adaptivity, as shown by the example demonstrated in Fig. 5 (with $K = 20$). Further performance improvement can be achieved by updating the covariance matrix using estimated signal sequentially such as described in (2). Info-Greedy Sensing also outperforms the sensing algorithm where a_i are chosen to be random Gaussian vectors with the same power allocation, as it uses prior knowledge (albeit being imprecise) about the signal distribution.

Fig. 6 demonstrates an effect that when there is a mismatch in the assumed covariance matrix, better performance can be achieved if we make many lower power measurements than making one full power measurement because we update the assumed covariance matrix in between.

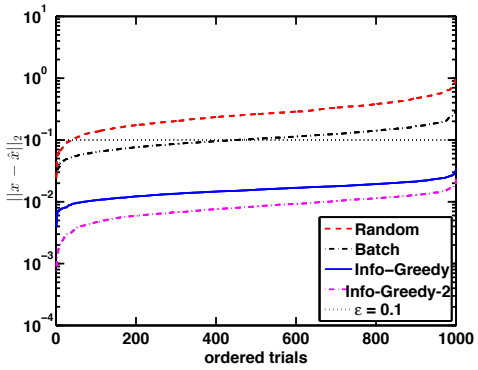


Fig. 5: Sensing a Gaussian signal of dimension $n = 100$, when there is mismatch between the assume covariance matrix and the true covariance matrix: $\widehat{\Sigma} \propto \Sigma + RR^\intercal$, where $R \in \mathbb{R}^{n\times 3}$ and each entry of $R_{ij} \sim \mathcal{N}(0,1)$. We repeat 1000 Monte Carlo trials and for each trial we use $K = 20$ measurements. The Info-Greedy-2 method corresponds to (2), where we update the assumed covariance matrix sequentially each time we recover a signal and $\alpha = 0.5$.

B. One-sparse measurements

In this example, we sense a GMM signal with a one-sparse measurement. Assume there are $C = 3$ components and we know the signal covariance matrix exactly. We consider two cases of generating the covariance matrix for each signal: when the low rank covariance matrices for each component are generated completely at random, and when it has certain structure. Fig. 7 shows the reconstruction error $\|\hat{x} - x\|$, using $K = 40$ one-sparse measurements for GMM signals. Note that Info-Greedy Sensing (Algorithm 3) with unit power $\beta_j = 1$ can significantly outperform the random approach with unit power (which corresponds to randomly selecting coordinates of the signal to measure). Fig. 8 also compares the mis-classification rate of Info-Greedy Sensing with one-sparse measurements to that with using a full signal vector x for classification. Note that, interestingly, using $K = 50$ one-sparse measurements we can obtain a performance very similar to the ideal case, which can be explained since we exploit the correlation structure of the signal.

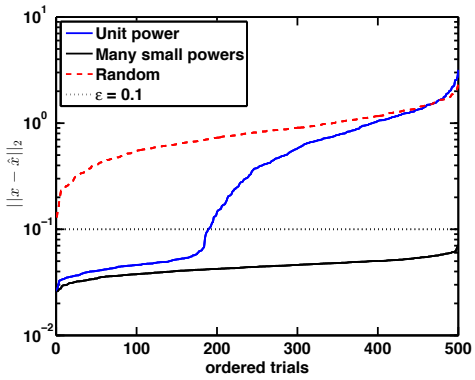
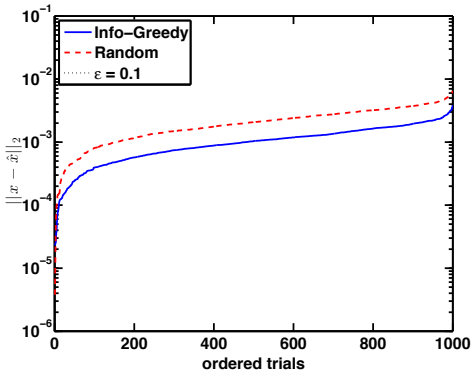
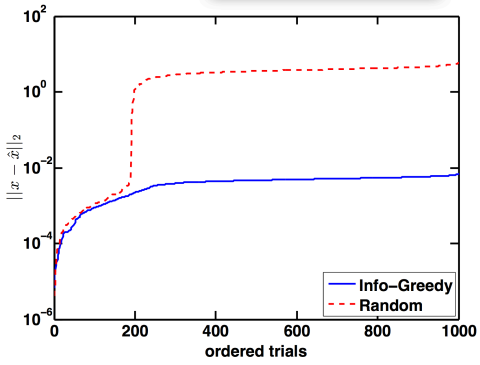


Fig. 6: Comparison of sensing a Gaussian signal with dimension $n = 100$ using unit power measurements along the eigenvector direction, versus splitting each unit-power measurement into 5 smaller ones, each with amplitude $\sqrt{1/5}$, and we update the covariance matrix in between. The mismatched covariance matrix is $\widehat{\Sigma} \propto \Sigma + rr^\intercal$, where $r \in \mathbb{R}^{n\times 5}$ and each entry of r is i.i.d. $\mathcal{N}(0,1)$, and $\widehat{\Sigma}$ is normalized to have unit spectral norm.



(a) $\Sigma_c \propto RR^\intercal$, $R \in \mathbb{R}^{n\times 3}$, $R_{ij} \sim \mathcal{N}(0,1)$



(b) $\Sigma_c \propto (11^\intercal + 20\alpha^2 \cdot \mathrm{diag}\{n, n-1, \cdots, 1\})$, $\alpha \sim \mathcal{N}(0,1)$

Fig. 7: Sensing a low-rank GMM signal of dimension $n = 100$ using $K = 40$ measurements with $\sigma = 0.001$, when the covariance matrices are generated (a) completely randomly, or (b) having certain structure. The covariance matrices Σ_c are normalized so that their spectral norms are 1.

C. Real data

1) Sensing of a video stream using Gaussian model: In this example, we use a video from the Solar Data Observatory. The frame is of size 232×292 pixels. We use the first 50 frames to form a sample covariance matrix $\widehat{\Sigma}$, and use it to perform Info-Greedy Sensing on the rest of the frames. We take $K = 90$

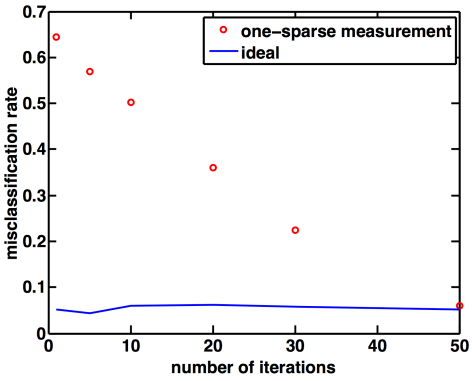


Fig. 8: Classifying a signal of dimension $n = 100$ generated from a GMM model with covariance matrix generated according to $\Sigma \propto RR^{\mathsf{T}}$ and the true distribution is $\pi = (0.5, 0.3, 0.2)$. We assume a uniform initial distribution $(1/3, 1/3, 1/3)$. Misclassification rate versus the number of measurements K. Ideal case corresponds to where we observe x and run a quadratic discriminate analysis using the full vector x (i.e. rather than just observing a noisy version of an entry of x each time).

measurements. As demonstrated in Fig. 9, Info-Greedy Sensing performs much better in that it acquires more information such that the recovered image has much richer details.

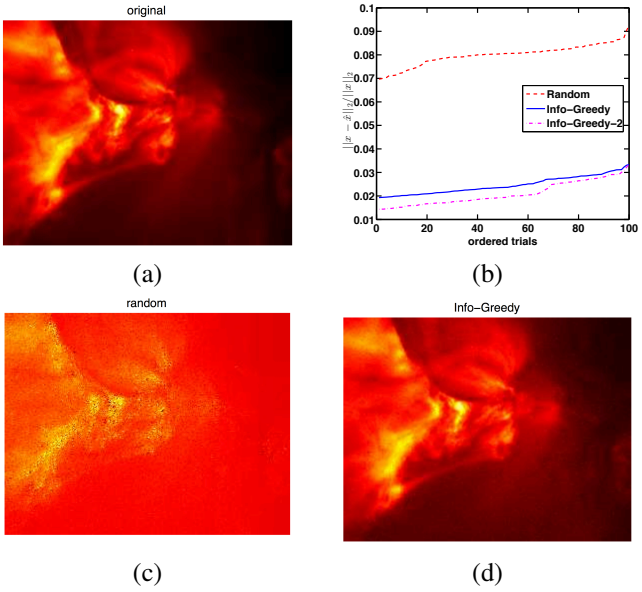


Fig. 9: Recovery of solar flare images of size 224 by 288 with $K = 90$ measurements and no sensing noise. We used the first 50 frames to estimate the mean and covariance matrix of a single Gaussian. (a) original image for 300th frame; (b) ordered relative recovery error of the 200th to the 300th frames; (c) recovered the 300th frame using random measurement; (d) recovered the 300th frame using Info-Greedy Sensing.

2) Sensing of a high-resolution image using GMM: We consider a scheme for sensing a high-resolution image that exploits the fact that the patches of the image can be approximated using a Gaussian mixture model, as demonstrated in Fig. 1. We break the image into 8 by 8 patches, which result in 89250 patches. We randomly select 500 patches (0.56% of the total pixels) to estimate a GMM model with $C = 10$ components, and then based on the estimated GMM initialize Info-Greedy Sensing with $K = 5$ measurements and sense the rest of the patches. This means we can use a compressive imaging system to capture 5 low resolution images of size 238-by-275 (this corresponds to compressing the data into 8.32% of its original dimensionality). With such a small number of measurements, the recovered image from Info-Greedy Sensing measurements has superior quality compared with those with random masks.

V. Conclusions and Discussions

In this paper, we have explored the value of information and how to use such information in sequential compressive sensing, by examining the Info-Greedy Sensing algorithms when the signal covariance matrix is not known exactly. We quantify the algorithm performances in the presence of estimation errors and when only one-sparse measurements are allowed.

Our results for Gaussian and GMM signals are quite general in the following sense. In high-dimensional problems, a commonly used low-dimensional signal model for x is to assume the signal lies in a subspace plus Gaussian noise, which corresponds to the case where the signal is Gaussian with a low-rank covariance matrix; GMM is also commonly used (e.g., in image analysis and video processing) as it models signals lying in a union of multiple subspaces plus Gaussian noise. In fact, parameterizing via low-rank GMMs is a popular way to approximate complex densities for high-dimensional data. Hence, we may be able to couple the results for Info-Greedy Sensing of GMM with the recently developed methods of scalable multi-scale density estimation based on empirical Bayes [29] to create powerful tools for information guided sensing for a general signal model. We may also be able to obtain performance guarantees using multiplicative weight update techniques together with the error bounds in [29].

References

[1] A. Ashok, P. Baheti, and M. A. Neifeld, "Compressive imaging system design using task-specific information," *Applied Optics*, vol. 47, no. 25, pp. 4457–4471, 2008.

[2] J. Ke, A. Ashok, and M. Neifeld, "Object reconstruction from adaptive compressive measurements in feature-specific imaging," *Applied Optics*, vol. 49, no. 34, pp. 27–39, 2010.

[3] A. Ashok and M. A. Neifeld, "Compressive imaging: hybrid measurement basis design," *J. Opt. Soc. Am. A*, vol. 28, no. 6, pp. 1041– 1050, 2011.

[4] W. Boonsong and W. Ismail, "Wireless monitoring of household electrical power meter using embedded RFID with wireless sensor network platform," *Int. J. Distributed Sensor Networks, Article ID 876914, 10 pages*, vol. 2014, 2014.

[5] B. Zhang, X. Cheng, N. Zhang, Y. Cui, Y. Li, and Q. Liang, "Sparse target counting and localization in sensor networks based on compressive sensing," in *IEEE Int. Conf. Computer Communications (INFOCOM)*, pp. 2255 – 2258, 2014.

[6] G. Braun, S. Pokutta, and Y. Xie, "Info-greedy sequential adaptive compressed sensing," *to appear in IEEE J. Sel. Top. Sig. Proc.*, 2014.

[7] J. Haupt, R. Nowak, and R. Castro, "Adaptive sensing for sparse signal recovery," in *IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop (DSP/SPE)*, pp. 702 – 707, 2009.
[8] M. A. Davenport and E. Arias-Castro, "Compressive binary search," *arXiv:1202.0937v2*, 2012.
[9] A. Tajer and H. V. Poor, "Quick search for rare events," *arXiv:1210:2406v1*, 2012.
[10] D. Malioutov, S. Sanghavi, and A. Willsky, "Sequential compressed sensing," *IEEE J. Sel. Topics Sig. Proc.*, vol. 4, pp. 435–444, April 2010.
[11] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak, "Sequentially designed compressed sensing," in *Proc. IEEE/SP Workshop on Statistical Signal Processing*, 2012.
[12] S. Jain, A. Soni, and J. Haupt, "Compressive measurement designs for estimating structured signals in structured clutter: A Bayesian experimental design approach," *arXiv:1311.5599v1*, 2013.
[13] M. L. Malloy and R. Nowak, "Near-optimal adaptive compressed sensing," *arXiv:1306.6239v1*, 2013.
[14] A. Krishnamurthy, J. Sharpnack, and A. Singh, "Recovering graph-structured activations using adaptive compressive measurements," in *Annual Asilomar Conference on Signals, Systems, and Computers*, Sept. 2013.
[15] T. Ervin and R. Castro, "Adaptive sensing for estimation of structure sparse signals," *arXiv:1311.7118*, 2013.
[16] S. Akshay and J. Haupt, "On the fundamental limits of recovering tree sparse vectors from noisy linear measurements," *IEEE Trans. Info. Theory*, vol. 60, no. 1, pp. 133–149, 2014.
[17] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Sig. Proc.*, vol. 56, no. 6, pp. 2346–2356, 2008.
[18] J. M. Duarte-Carvajalino, G. Yu, L. Carin, and G. Sapiro, "Task-driven adaptive statistical compressive sensing of Gaussian mixture models," *IEEE Trans. Sig. Proc.*, vol. 61, no. 3, pp. 585–600, 2013.
[19] W. Carson, M. Chen, R. Calderbank, and L. Carin, "Communication inspired projection design with application to compressive sensing," *SIAM J. Imaging Sciences*, 2012.
[20] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. Nowak, "Covariance sketching," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, 2012.
[21] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. Nowak, "Sketching sparse matrices," *ArXiv ID:1303.6544*, 2013.
[22] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Trans. Info. Theory*, vol. in revision, 2013.
[23] C. Hellier, *Handbook of Nondestructive Evaluation*. McGraw-Hill, 2003.
[24] D. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Info. Theory*, pp. 141–154, 2006.
[25] M. Payaró and D. P. Palomar, "Hessian and concavity of mutual information, entropy, and entropy power in linear vector Gaussian channels," *IEEE Trans. Info. Theory*, pp. 3613–3628, Aug. 2009.
[26] N. Cesa-Bianchi, G. Lugosi, *et al.*, *Prediction, learning, and games*, vol. 1. Cambridge University Press Cambridge, 2006.
[27] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic game theory*. Cambridge University Press, 2007.
[28] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: A meta-algorithm and applications," *Theory of Computing*, vol. 8, no. 1, pp. 121–164, 2012.
[29] Y. Wang, A. Canale, and D. Dunson, "Scalable multiscale density estimation," *arXiv:1410.7692*, 2014.
[30] G. W. Stewart and J.-G. Sun, *Matrix perturbation theory*. Academic Press, Inc., 1990.
[31] S. Zhu, "A short note on the tail bound of wishart distribution," *arXiv:1212.5860*, 2012.
[32] T. Vincent, L. Tenorio, and M. Wakin, *Concentration of measure: fundamentals and tools.* Rice University, Lecture Notes.

APPENDIX A
COVARIANCE SKETCHING

Consider the following setup for covariance sketching. Suppose we are able to form measurement in the form of $y = a^\intercal x + w$ like we have in the Info-Greedy Sensing algorithm. Suppose there are N copies of Gaussian signal we would like to sketch: $\tilde{x}_1, \ldots, \tilde{x}_N$ that are i.i.d. sampled from $\mathcal{N}(0, \Sigma)$, and we sketch using M random vectors: $b_1, \ldots, b_M$. Then for each fixed sketching vector b_i, and fixed copy of the signal $\tilde{x}_j$, we acquire L noisy realizations of the projection result y_{ijl} via

$$y_{ijl} = b_i^\intercal \tilde{x}_j + w_{ijl}, \quad l = 1, \ldots, L.$$

We choose the random sampling vectors b_i as i.i.d. Gaussian with zero mean and covariance matrix equal to an identity matrix. Then we average y_{ijl} over all realizations $l = 1, \ldots, L$ to form the ith sketch y_{ij} for a single copy $\tilde{x}_j$:

$$y_{ij} = b_i^\intercal \tilde{x}_j + \underbrace{\frac{1}{L}\sum_{l=1}^{L} w_{ijl}}_{w_{ij}}.$$

The average is introduced to suppress measurement noise, which can be viewed as a generalization of sketching using just one sample. Denote $w_{ij} = \frac{1}{L}\sum_{l=1}^{L} w_{ijl}$, which is distributed as $\mathcal{N}(0, \sigma^2/L)$. Then we will use the average energy of the sketches as our data γ_i, $i = 1, \ldots, M$, for covariance recovery:

$$\gamma_i \triangleq \frac{1}{N}\sum_{j=1}^{N} y_{ij}^2.$$

Note that γ_i can be further expanded as

$$\gamma_i = \mathrm{tr}(\widehat{\Sigma}_N b_i b_i^\intercal) + \frac{2}{N}\sum_{j=1}^{N} w_{ij} b_i^\intercal \tilde{x}_j + \frac{1}{N}\sum_{j=1}^{N} w_{ij}^2, \qquad (16)$$

where

$$\widehat{\Sigma}_N = \frac{1}{N}\sum_{j=1}^{N} \tilde{x}_j \tilde{x}_j^\intercal,$$

is the maximum likelihood estimate of Σ (and is also unbiased). We can write (16) in vector matrix notation as follows. Let $\gamma = [\gamma_1, \cdots \gamma_M]^\intercal$. Define a linear operator $\mathcal{B} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^M$ such that $[B(X)]_i = \mathrm{tr}(X b_i b_i^\intercal)$. Thus, we can write (16) as a linear measurement of the true covariance matrix Σ

$$\gamma = \mathcal{B}(\Sigma) + \eta,$$

where $\eta \in \mathbb{R}^M$ contains all the error terms and corresponds to the noise in our covariance sketching measurements, with the ith entry given by

$$\eta_i = b_i^\intercal(\widehat{\Sigma}_N - \Sigma) b_i + \frac{2}{N}\sum_{j=1}^{N} w_{ij} b_i^\intercal \tilde{x}_j + \frac{1}{N}\sum_{j=1}^{N} w_{ij}^2.$$

Note that we can further bound the ℓ_1 norm of the error term as

$$\|\eta\|_1 = \sum_{i=1}^{M} |\eta_i| \le \|\widehat{\Sigma}_N - \Sigma\| b + 2\sum_{i=1}^{M} |z_i| + w,$$

where

$$b = \sum_{i=1}^{M} \|b_i\|^2, \ \mathbb{E}[b] = Mn, \ \mathrm{Var}[b] = 2Mn,$$

$$w = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij}^2,\ \mathbb{E}[w] = M\sigma^2/L,\ \text{and}\ \mathrm{Var}[w] = \frac{2M\sigma^4}{NL^2},$$

$$z_i = \frac{1}{N}\sum_{j=1}^{N} w_{ij} b_i^{\intercal}\tilde{x}_j,\ \mathbb{E}[z_i] = 0\ \text{and}\ \mathrm{Var}[z_i] = \frac{\sigma^2 \mathrm{tr}(\Sigma)}{NL}.$$

We may recover the true covariance matrix from the sketches γ using the convex optimization problem (9).

APPENDIX B
BACKGROUNDS

Lemma 4 (Eigenvalue of perturbed matrix [30])**.** *Let Σ, $\widehat{\Sigma} \in \mathbb{R}^{n\times n}$ be symmetric,with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$, respectively. Let $E = \widehat{\Sigma} - \Sigma$ have eigenvalues $e_1 \geq \cdots \geq e_n$. Then for each $i \in \{1, \cdots, n\}$, the perturbed eigenvalues satisfy $\hat{\lambda}_i \in [\lambda_i + e_n, \lambda_i + e_1]$.*

Lemma 5 (Stability conditions for covariance sketching [22])**.** *Denote $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ a linear operator and for $X \in \mathbb{R}^{n\times n}$, $\mathcal{A}(X) = \{a_i^T X a_i\}_{i=1}^m$. Suppose the measurement is contaminated by noise $\eta \in R^m$, i.e., $Y = \mathcal{A}(\Sigma) + \eta$ and assume $\|\eta\|_1 \leq \epsilon_1$. Then with probability exceeding $1 - \exp(-c_1 m)$ the solution $\widehat{\Sigma}$ to the trace minimization (9) satisfies*

$$\|\widehat{\Sigma} - \Sigma\|_F \leq c_0 \frac{\|\Sigma - \Sigma_r\|_*}{\sqrt{r}} + c_2 \frac{\epsilon_1}{m},$$

for all $\Sigma \in R^{n\times n}$, provided that $m > c_0 nr$. Here c_0, c_1, and c_2 are absolute constants and Σ_r represents the best rank-r approximation of Σ. When Σ_r is exactly rank-r

$$\|\widehat{\Sigma} - \Sigma\|_F \leq c_0 \frac{\epsilon_1}{m}.$$

Lemma 6 (Concentration of measure for Wishart distribution [31])**.** *If $X \in \mathbb{R}^{n\times n} \sim \mathcal{W}_n(N, \Sigma)$, then for $t > 0$,*

$$P\{\|\frac{1}{N}X - \Sigma\| \geq (\sqrt{\frac{2t(\theta+1)}{N}} + \frac{2t\theta}{N})\|\Sigma\|\} \leq 2n\exp(-t),$$

where $\theta = \mathrm{tr}(\Sigma)/\|\Sigma\|$.

APPENDIX C
PROOFS

A. Gaussian signal, with mismatch

Proof of Theorem 1. Let $\xi_k = \hat{\mu}_k - \mu_k$. From the update equation for the mean $\hat{\mu}_k = \hat{\mu}_{k-1} + \widehat{\Sigma}_{k-1}a_k(y_k - a_k^\intercal\hat{\mu}_{k-1})/(\hat{a}_k^\intercal\widehat{\Sigma}_{k-1}a_k + \sigma^2)$, since a_k is eigenvector of $\widehat{\Sigma}_{k-1}$, we have the following recursion:

$$\begin{aligned}\xi_k = (I_n - \frac{\hat{\lambda}_k a_k a_k^\intercal}{\beta_k\hat{\lambda}_k + \sigma^2})\xi_{k-1} \\ + \left[-\hat{\lambda}_k\frac{a_k^\intercal E_{k-1}a_k}{(\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k)(\beta_k\hat{\lambda}_k + \sigma^2)}a_k\right. \\ \left. + \frac{E_{k-1}a_k}{\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k}\right](a_k^\intercal(x - \mu_{k-1}) + w_k).\end{aligned} \tag{17}$$

From the recursion of ξ_k in (17), for some vector C_k defined properly, we have that

$$\begin{aligned}\mathbb{E}[\xi_k] = (I - \frac{\hat{\lambda}_k\beta_k}{\beta_k\hat{\lambda}_k + \sigma^2}u_k u_k^\intercal)\mathbb{E}[\xi_{k-1}] \\ + C_k\underbrace{\mathbb{E}[a_k^\intercal(x - \mu_{k-1}) + w_k]}_{0}.\end{aligned} \tag{18}$$

Note that the second term is equal to zero using an argument based on iterated expectation

$$\mathbb{E}[a_k^\intercal(x - \mu_{k-1}) + w_k] = a_k^\intercal\mathbb{E}[\mathbb{E}[x - \mu_{k-1}|y_1, \ldots, y_k]] = 0.$$

Hence Theorem 1 is proved by iteratively apply the recursion (18). When $\hat{\mu}_0 - \mu_0 = 0$, we have $\mathbb{E}[\xi_k] = 0, k = 0, 1, \ldots, K$. □

In the following, Lemma 7 to Lemma 9 are used to prove Theorem 2.

Lemma 7 (Recursion in covariance matrix mismatch.)**.** *If $\delta_{k-1} \le 3\sigma^2/4\beta_k$, then $\delta_k \le 4\delta_{k-1}$.*

Proof. Let $\widehat{A}_k = a_k a_k^\intercal$. Hence, $\|\widehat{A}_k\| = \beta_k$. Recall that a_k is the eigenvector of $\widehat{\Sigma}_{k-1}$, using the definition of $E_k = \widehat{\Sigma}_k - \Sigma_k$, together with the recursions of the covariance matrices

$$\widehat{\Sigma}_k = \widehat{\Sigma}_{k-1} - \widehat{\Sigma}_{k-1}a_k a_k^\intercal\Sigma_{k-1}/(\hat{\lambda}_k + \sigma^2), \tag{19}$$

$$\Sigma_k = \Sigma_{k-1} - \Sigma_{k-1}a_k a_k^\intercal\Sigma_{k-1}/(a_k^\intercal\Sigma_{k-1}a_k + \sigma^2), \tag{20}$$

we have

$$E_k = E_{k-1} + \frac{\Sigma_{k-1}a_k a_k^\intercal\Sigma_{k-1}}{a_k^\intercal\Sigma_{k-1}a_k + \sigma^2} - \frac{\hat{\lambda}_k a_k a_k^\intercal\widehat{\Sigma}_{k-1}}{\beta_k\hat{\lambda}_k + \sigma^2}.$$

Based on this recursion, using $\delta_k = \|E_k\|$, the triangle inequality, and Cauchy-Schwartz inequality $\|AB\| \le \|A\|\|B\|$, we have

$$\begin{aligned}\delta_k &\le \delta_{k-1} + \frac{\beta_k\hat{\lambda}_k a_k E_{k-1}a_k}{(\beta_k\hat{\lambda}_k + \sigma^2)(\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a)}\cdot\|\widehat{A}_k\widehat{\Sigma}_{k-1}\| \\ &\quad + \frac{1}{\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k} \\ &\qquad \cdot[\hat{\lambda}_k(\|\widehat{A}_k E_{k-1}\| + \|E_{k-1}\widehat{A}_k\|) + \|E_{k-1}\widehat{A}_k E_{k-1}\|] \\ &\le \delta_{k-1} + \frac{\beta_k^2\hat{\lambda}_k^2\delta_{k-1}}{(\beta_k\hat{\lambda}_k + \sigma^2)(\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1})} \\ &\quad + \frac{\beta_k}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}[2\hat{\lambda}_k\delta_{k-1} + \delta_{k-1}^2] \\ &\le (1 + \frac{3\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}})\delta_{k-1} \\ &\quad + \frac{\beta_k}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}\delta_{k-1}^2.\end{aligned}$$

Hence, if set $\delta_{k-1} \le 3\sigma^2/(4\beta_k)$, i.e., $\delta_{k-1}\beta_k \le \frac{3}{4}\sigma^2$, the last inequality can be upper bounded by

$$(1 + 3\cdot\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k + \sigma^2/4})\delta_{k-1} + 3\cdot\frac{\sigma^2/4}{\beta_k\hat{\lambda}_k + \sigma^2/4}\delta_{k-1} = 4\delta_{k-1}.$$

Hence, if $\delta_{k-1} \le 3\sigma^2/(4\beta_k)$, we have $\delta_k \le 4\delta_{k-1}$. □

Lemma 8 (Recursion for trace of the true covariance matrix)**.** *If $\delta_{k-1} \le \hat{\lambda}_k$,*

$$\mathrm{tr}(\Sigma_k) \le \mathrm{tr}(\Sigma_{k-1}) - \frac{\beta_k\hat{\lambda}_k^2}{\beta_k\hat{\lambda}_k + \sigma^2} + \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k + \sigma^2 - \beta_k\delta_{k-1}}. \tag{21}$$

Proof. Let $\widehat{A}_k = a_k a_k^\intercal$. Using the definition of E_k and the recursions (19) and (20), the perturbation matrix E_k after k iterations is given by

$$\begin{aligned}E_k = E_{k-1} + \hat{\lambda}_k^2\widehat{A}_k\cdot\frac{a_k^\intercal E_{k-1}a_k}{(\beta_k\hat{\lambda}_k + \sigma^2)(\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k)} \\ - \frac{\hat{\lambda}_k}{\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k}\cdot(\widehat{A}_k E_{k-1} + E_{k-1}\widehat{A}_k) \\ + \frac{1}{\beta_k\hat{\lambda}_k + \sigma^2 - a_k^\intercal E_{k-1}a_k}E_{k-1}\widehat{A}_k E_{k-1}.\end{aligned} \tag{22}$$

Note that $\mathrm{rank}(\widehat{A}_k) = 1$, thus $\mathrm{rank}(\widehat{A}_k E_{k-1}) \le 1$, therefore it has at most one nonzero eigenvalue,

$$\begin{aligned}|\mathrm{tr}(\widehat{A}_k E_{k-1})| &= |\mathrm{tr}(E_{k-1}\widehat{A}_k)| \\ &= \|\widehat{A}_k E_{k-1}\| \le \|\widehat{A}_k\|\|E_{k-1}\| = \beta_k\delta_{k-1}.\end{aligned}$$

Note that E_{k-1} is symmetric and $\hat{A}_k$ is positive semi-definite,

we have $\mathrm{tr}(E_{k-1}\widehat{A}_k E_{k-1}) \geq 0$. Hence, from (22) we have

$$\begin{aligned}\mathrm{tr}(E_k) &= \mathrm{tr}(\widehat{\Sigma}_\mathrm{k}) - \mathrm{tr}(\Sigma_\mathrm{k}) \\ &\geq \mathrm{tr}(E_{k-1}) - \frac{3\beta_k\hat{\lambda}_k(\beta_k\hat{\lambda}_k + \frac{2\sigma^2}{3})\delta_{k-1}}{(\beta_k\hat{\lambda}_k+\sigma^2)(\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1})} \\ &\geq \mathrm{tr}(E_{k-1}) - \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}}.\end{aligned}$$

After rearranging terms we obtain

$$\mathrm{tr}(\Sigma_k) \leq \mathrm{tr}(\Sigma_{k-1}) + [\mathrm{tr}(\widehat{\Sigma}_k) - \mathrm{tr}(\widehat{\Sigma}_{k-1})] + \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}}.$$

Together with the recursion for trace of $\mathrm{tr}(\widehat{\Sigma}_k)$ in (6), we have

$$\mathrm{tr}(\Sigma_k) \leq \mathrm{tr}(\Sigma_{k-1}) - \frac{\beta_k\hat{\lambda}_k^2}{\beta_k\hat{\lambda}_k+\sigma^2} + \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}}.$$

□

Lemma 9. *For a given positive semi-definite matrix $X \in \mathbb{R}^{n\times n}$, and a vector $h \in \mathbb{R}^n$, if*

$$Y = X - \frac{1}{h^\intercal X h + \sigma^2} X h h^\intercal X,$$

then $\mathrm{rank}(X) = \mathrm{rank}(Y)$.

Proof. Apparently, for all $x \in \ker(X)$, $Yx = 0$, i.e., $\ker(X) \subset \ker(Y)$. Decompose $X = Q^\intercal Q$. For all $x \in \ker(Y)$, let $b = Qh$, $z = Qx$. If $b = 0$, $Y = X$; otherwise, when $b \neq 0$, we have

$$0 = x^\intercal Y x = z^\intercal z - \frac{z^\intercal b b^\intercal z}{b^\intercal b + \sigma^2}.$$

Thus,

$$z^\intercal z = \frac{z^\intercal b b^\intercal z}{b^\intercal b+\sigma^2} \leq \frac{b^\intercal b}{b^\intercal b+\sigma^2} z^\intercal z.$$

Therefore $z = 0$, i.e. $x \in \ker(X)$, $\ker(Y) \subset \ker(X)$. This shows that $\ker(X) = \ker(Y)$, or equivalently $\mathrm{rank}(X) = \mathrm{rank}(Y)$. □

Proof of Theorem 2. Recall that for $k = 1, \ldots, K$, $\hat{\lambda}_k \geq \chi_{n,p,\varepsilon}$. Using Lemma 7, we can show that for some $0 < \delta < 1$, if $\delta_0 \leq \delta\chi_{n,p,\varepsilon}/4^{K+1} \leq 3\sigma^2/(4^{K+1}\beta_1)$, (the second inequality comes from the fact that $(1/\chi_{n,p,\varepsilon} - 1/\hat{\lambda}_1)\chi_{n,p,\varepsilon}\sigma^2 \leq 3\sigma^2$), then for the first K measurements,

$$\delta_k \leq \frac{1}{4^{K-k+1}}\frac{\delta\chi_{n,p,\varepsilon}}{4} \leq \frac{1}{4^{K-k}}\frac{3\sigma^2}{4\beta_1}, \quad k = 1, \ldots, K.$$

Clearly,

$$\delta_{k-1} \leq \delta\chi_{n,p,\varepsilon}/16.$$

Hence,

$$(4+\delta)\delta_{k-1} \leq \delta\lambda_k.$$

Note that $\beta_k\delta_{k-1} \leq \sigma^2$ and $|\lambda_k - \hat{\lambda}_k| \leq \delta_{k-1}$, we have

$$\beta_k\lambda_k \leq \beta_k(\hat{\lambda}_k + \delta_{k-1}) \leq \beta_k\hat{\lambda}_k + \sigma^2.$$

Thus,

$$4\delta_{k-1}(\beta_k\hat{\lambda}_k+\sigma^2) + \delta\beta_k\lambda_k\delta_{k-1} \leq \delta\lambda_k(\beta_k\hat{\lambda}_k+\sigma^2).$$

Then we have

$$3\beta_k\hat{\lambda}_k\delta_{k-1}(\beta_k\hat{\lambda}_k+\sigma^2) \leq \beta_k\hat{\lambda}_k(\delta\lambda_k-\delta_{k-1})(\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}),$$

which can be rewritten as

$$\frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}} \leq \frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}(\delta\lambda_k - \delta_{k-1}).$$

Hence,

$$\frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}} \leq \frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}[(\delta-1)\lambda_k + \hat{\lambda}_k],$$

which can be written as

$$-\frac{\beta_k\hat{\lambda}_k^2}{\beta_k\hat{\lambda}_k+\sigma^2} + \frac{3\beta_k\hat{\lambda}_k\delta_{k-1}}{\beta_k\hat{\lambda}_k+\sigma^2-\beta_k\delta_{k-1}} \leq -(1-\delta)\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}\lambda_k.$$

By applying Lemma 8, we have

$$\begin{aligned}\mathrm{tr}(\Sigma_k) &\leq \mathrm{tr}(\Sigma_{k-1}) - (1-\delta)\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}\lambda_k \\ &\leq \mathrm{tr}(\Sigma_{k-1}) - (1-\delta)\frac{\beta_k\hat{\lambda}_k}{\beta_k\hat{\lambda}_k+\sigma^2}\frac{\mathrm{tr}(\Sigma_{k-1})}{s} \triangleq f_k\mathrm{tr}(\Sigma_{k-1}),\end{aligned}$$

where we have used the definition for f_k in (4). Subsequently,

$$\mathrm{tr}(\Sigma_k) \leq (\prod_{j=1}^{k} f_j)\mathrm{tr}(\Sigma_0).$$

Lemma 9 shows that the rank of the covariance will not be changed by updating the covariance matrix sequentially: $\mathrm{rank}(\Sigma_1) = \cdots = \mathrm{rank}(\Sigma_k) = s$. Hence, we may decompose the covariance matrix $\Sigma_k = QQ^\intercal$, with $Q \in \mathbb{R}^{n\times s}$ being a full-rank matrix, then $\mathsf{Vol}(\Sigma_k) = \det(Q^\intercal Q)$. Since $\mathrm{tr}(Q^\intercal Q) = \mathrm{tr}(QQ^\intercal)$, we have

$$\begin{aligned}\mathsf{Vol}^2(\Sigma_k) = \det(Q^\intercal Q) &\overset{(1)}{\leq} \prod_{j=1}^{s}(Q^\intercal Q)_{jj} \\ &\overset{(2)}{\leq} \left(\frac{\mathrm{tr}(Q^\intercal Q)}{s}\right)^s = \left(\frac{\mathrm{tr}(\Sigma_\mathrm{k})}{s}\right)^s,\end{aligned}$$

where (1) follows from the Hadamard's inequality and (2) follows from the mean inequality. Finally, we can bound the conditional entropy of the signal as

$$\begin{aligned}\mathbb{H}\left[x \,|\, y_j, a_j, j \leq k\right] &= \ln(2\pi e)^{s/2}\mathsf{Vol}(\Sigma_k) \\ &\leq \frac{s}{2}\ln\{2\pi e(\prod_{j=1}^{k} f_j)\mathrm{tr}(\Sigma_0)\},\end{aligned}$$

which leads to the desired result. □

Proof of Theorem 3. Recall that $\mathrm{rank}(\Sigma) = s$, and hence $\lambda_k = 0$, $k = s+1, \ldots, n$. Note that for each iteration, the eigenvalue of $\widehat{\Sigma}_k$ in the direction of a_k, which corresponds to

the largest eigenvalue of $\widehat{\Sigma}_k$, is eliminated below the threshold $\chi_{n,p,\varepsilon}$. Therefore, as long as the algorithm continues, the largest eigenvalue of $\widehat{\Sigma}_k$ is exactly the $(k+1)$th largest eigenvalue of $\widehat{\Sigma}$. Now if

$$\delta_0 \le \chi_{n,p,\varepsilon}/4^{s+1}, \tag{23}$$

using Lemma 4 and Lemma 7, we have that

$$|\hat{\lambda}_k - \lambda_k| \le \delta_0, \text{ for } k = 1, \dots, s,$$
$$|\hat{\lambda}_j| \le \delta_0 \le \chi_{n,p,\varepsilon} - \delta_s, \text{ for } k = s+1, \dots, n.$$

In the ideal case without perturbation, each measurement decreases the eigenvalue along a given eigenvector to be below $\chi_{n,p,\varepsilon}$. Suppose in the ideal case, the algorithm terminates at $K \le s$ iterations, which means

$$\lambda_1 \ge \cdots \ge \lambda_L \ge \chi_{n,p,\varepsilon} > \lambda_{K+1}(\Sigma) \ge \cdots \ge \lambda_s(\Sigma),$$

and the total power needed is

$$P_{\text{ideal}} = \sum_{k=1}^{K} \sigma^2 \left(\frac{1}{\chi_{n,p,\varepsilon}} - \frac{1}{\lambda_k} \right). \tag{24}$$

On the other hand, in the presence of perturbation, the algorithm will terminate using more than K iterations since with perturbation, eigenvalues of Σ that originally below $\chi_{n,p,\varepsilon}$ may get above $\chi_{n,p,\varepsilon}$. In this case, we will also allocate power while taking into account the perturbation:

$$\beta_k = \sigma^2 \left(\frac{1}{\chi_{n,p,\varepsilon} - \delta_s} - \frac{1}{\hat{\lambda}_k} \right).$$

This suffices to eliminate even the smallest eigenvalue to be below threshold $\chi_{n,p,\varepsilon}$ since

$$\frac{\sigma^2 \hat{\lambda}_{k-1}}{\beta_{k-1}\hat{\lambda}_{k-1} + \sigma^2} = \chi_{n,p,\varepsilon} - \delta_s < \chi_{n,p,\varepsilon}.$$

We first estimate the total amount of power used at most to eliminate eigenvalues $\hat{\lambda}_k$, for $K+1 \le k \le s$:

$$\begin{aligned}
\beta_k &= \sigma^2(1/(\chi_{n,p,\varepsilon} - \delta_s) - 1/\hat{\lambda}_k) \\
&\le \sigma^2(1/(\chi_{n,p,\varepsilon} - \delta_s) - 1/(\chi_{n,p,\varepsilon} + \delta_0)) \\
&\le \sigma^2 \frac{(4^s+1)\delta_0}{(\chi_{n,p,\varepsilon} - 4^s\delta_0)(\chi_{n,p,\varepsilon} + \delta_0)} \le \frac{20}{51}\frac{\sigma^2}{\chi_{n,p,\varepsilon}}.
\end{aligned}$$

where we have used the fact that $\delta_s \le 4^s\delta_0$ (a consequence of Lemma 7), the assumption (23), and monotonicity of the upper bound in s. The total power to reach precision ε in the presence of mismatch can be upper bounded by

$$\begin{aligned}
P_{\text{mismatch}} &\le \sum_{k=1}^{s} \beta_k \\
&\le \sigma^2 \left\{ \sum_{k=1}^{K} \left(\frac{1}{\chi_{n,p,\varepsilon} - \delta_s} - \frac{1}{\hat{\lambda}_k} \right) + \frac{20(s-K)}{51}\frac{\sigma^2}{\chi_{n,p,\varepsilon}} \right\}.
\end{aligned}$$

In order to achieve precision ε and confidence level p, the extra power needed is upper bounded as

$$\begin{aligned}
&P_{\text{mismatch}} - P_{\text{ideal}} \\
&\le \sigma^2 \left\{ \sum_{k=1}^{K} \left(\frac{1}{3}\frac{1}{\chi_{n,p,\varepsilon}} + \frac{\delta_0}{\lambda_k^2} \right) + \frac{20(s-K)}{51}\frac{1}{\chi_{n,p,\varepsilon}} \right\} \\
&\le \sigma^2 \left\{ \frac{1}{4^{s+1}} \sum_{k=1}^{K} \frac{\chi_{n,p,\varepsilon}}{\lambda_k^2} + \frac{20s-3K}{51}\frac{1}{\chi_{n,p,\varepsilon}} \right\} \\
&< \left(\frac{20}{51}s - (\frac{3}{51} - \frac{1}{4^{s+1}})K \right) \frac{\sigma^2}{\chi_{n,p,\varepsilon}} \\
&\le \left(\frac{20}{51}s + \frac{1}{272}K \right) \frac{\sigma^2}{\chi_{n,p,\varepsilon}},
\end{aligned}$$

where we have again used $\delta_s \le 4^s\delta_0 \le 4^s\chi_{n,p,\varepsilon}/4^{s+1} = \chi_{n,p,\varepsilon}/4$, $1/\hat{\lambda}_k - 1/\lambda_k \le \delta_0/\lambda_k^2$, the fact that $\lambda_k \ge \chi_{n,p,\varepsilon}$ for $k = 1, \dots, K$. □

Proof of Lemma 1. It is a direct consequence of Lemma 6. Let $\theta = \text{tr}(\Sigma)/\|\Sigma\| \ge 1$. For some constant $\delta > 0$, set

$$L \ge 4n^{1/2}\text{tr}(\Sigma)(\|\Sigma\|/\delta^2 + 4/\delta).$$

Then from Lemma 6, we have

$$\begin{aligned}
&P\{\|\widehat{\Sigma} - \Sigma\| \le \delta\} \\
&\ge P\{\|\widehat{\Sigma} - \Sigma\| \le \left(\sqrt{2n^{1/2}(\theta+1)/L} + 2\theta n^{1/2}/L \right) \|\Sigma\|\} \\
&> 1 - 2n\exp(-\sqrt{n}).
\end{aligned}$$

□

The following Lemma is used in the proof of Lemma 2.

Lemma 10. *For the setup in Section A, if for some constant M, N and L satisfies the conditions in Lemma 2, then $\|\eta\|_1 \le \tau$ with probability exceeding $1 - 2/n - 2/\sqrt{n} - 2n\exp(-c_1 M)$ for some universal constant $c_1 > 0$.*

Proof. Let $\theta = \text{tr}(\Sigma)/\|\Sigma\|$. From Chebyshev's inequality, we have that

$$\mathbb{P}\{|z_i| < \frac{\tau}{6M}\} \ge 1 - \frac{36M^2\sigma^2\text{tr}(\Sigma)}{NL\tau^2}, \quad i = 1, \dots, K$$
$$\mathbb{P}\{|w| < M\frac{\sigma^2}{L} + \frac{\tau}{6}\} \ge 1 - \frac{72\sigma^4 M}{NL^2\tau^2},$$

and

$$\mathbb{P}\{|b| < (M + \sqrt{M})n\} \ge 1 - \frac{2}{n}.$$

When

$$N \ge 4n^{1/2}\text{tr}(\Sigma)(\frac{36n^2M^2\|\Sigma\|}{\tau^2} + \frac{24nM}{\tau}), \tag{25}$$

with the concentration inequality for Wishart distribution in Lemma 6 and plugging in the lower bound for N in (25) and

the definition for τ in (13) we have

$$\begin{aligned}&\mathbb{P}\{\|\widehat{\Sigma}_N-\Sigma\|\le\tau/[3n(M+\sqrt{M})]\}\\ &\ge\mathbb{P}\{\|\widehat{\Sigma}_N-\Sigma\|\le(\sqrt{\frac{2n^{1/2}\theta}{N}}+\frac{2\theta n^{1/2}}{N})\|\Sigma\|\}\\ &>1-2n\exp(-\sqrt{n}).\end{aligned}$$

Furthermore, when L satisfies (12), we have

$$\begin{aligned}&\mathbb{P}\{|z_i|<\frac{\tau}{6M}\}\ge1-\frac{1}{M\sqrt{n}},\\ &\mathbb{P}\{|w|<\frac{\tau}{3}\}\ge1-\frac{1}{\sqrt{n}},\\ &\mathbb{P}\{|b|<(M+\sqrt{M})n\}\ge1-\frac{2}{n}.\end{aligned}$$

Therefore, $\|\eta\|_1\le\tau$ holds with probability at least $1-2/n-2/\sqrt{n}-2n\exp(-\sqrt{n})$. □

Proof of Lemma 2. With Lemma 10, let $\tau=M\delta/c_2$, the choices of M, N, and L ensure that $\|\eta\|_1\le M\delta/c_2$ with probability at least $1-2/n-2/\sqrt{n}-2n\exp(-\sqrt{n})$. By Lemma 5 and noting that the rank of Σ is s, we have $\|\widehat{\Sigma}-\Sigma\|_F\le\delta$. Therefore, with probability exceeding $1-2/n-2/\sqrt{n}-2n\exp(-\sqrt{n})-\exp(-c_0c_1ns)$,

$$\|\widehat{\Sigma}-\Sigma\|\le\|\widehat{\Sigma}-\Sigma\|_F\le\delta.$$

□

The proof of Theorem 4 will use the following two lemmas.

Lemma 11 (Moment generating function of multivariate Gaussian [32])**.** *Assume $X\sim\mathcal{N}(0,\Sigma)$. The moment generating function of $\|X\|_2$ is*

$$\mathbb{E}[e^{s\|X\|_2}]=1/\sqrt{I-2s\Sigma}.$$

Proof of Theorem 4. We adapt the technique used in [6] for proving performance bound for GMM signal without mismatch. Suppose the true signal is generated from the c^*th component. First, apply measurements to each component $c\in[C]$. Clearly, spending a total amount of power $\sum_{c=1}^{C}m_c$ would suffice to ensure that the norm of covariance of each individual component is below $\chi_{n,p,\varepsilon}$. In the ideal case, the weight is updated in the following manner:

$$\pi_c^{k+1}=\pi_c^kL_k\exp\{-\frac{1}{2}\frac{(y_k-a_k^\intercal\mu_{c,k-1})^2}{a_k^\intercal\Sigma_{c,k-1}a_k+\sigma^2}\},\quad c\in[C].$$

In the presence of mismatch, this becomes

$$\hat{\pi}_c^{k+1}=\hat{\pi}_c^k\hat{L}_k\exp\{-\frac{1}{2}\frac{(y_k-a_k^\intercal\hat{\mu}_{c,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{c,k-1}a_k+\sigma^2}\},\quad c\in[C].$$

The L_k and $\hat{L}_k$ for $k=1,2,\dots$ are normalization coefficients. After m measurements,

$$\begin{aligned}&\frac{1}{m}\left|\sum_{k=1}^{m}\sum_{\ell=1}^{C}\frac{(y_k-a_k^\intercal\hat{\mu}_{\ell,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{\ell,k-1}a_k+\sigma^2}\cdot\hat{\pi}_\ell^k-\frac{(y_k-a_k^\intercal\mu_{c^*,k-1})^2}{a_k^\intercal\Sigma_{c^*,k-1}a_k+\sigma^2}\right|\\ &=\frac{1}{m}\left|\sum_{k=1}^{m}\sum_{\ell=1}^{C}\frac{(y_k-a_k^\intercal\hat{\mu}_{\ell,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{\ell,k-1}a_k+\sigma^2}\cdot\hat{\pi}_\ell^k-\sum_{k=1}^{m}\frac{(y_k-a_k^\intercal\hat{\mu}_{c^*,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{c^*,k-1}a_k+\sigma^2}\right.\\ &\qquad\left.+\sum_{k=1}^{m}(\frac{(y_k-a_k^\intercal\hat{\mu}_{c^*,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{c^*,k-1}a_k+\sigma^2}-\frac{(y_k-a_k^\intercal\mu_{c^*,k-1})^2}{a_k^\intercal\Sigma_{c^*,k-1}a_k+\sigma^2})\right|\\ &\le\hat{\eta}+\frac{2\ln|C|}{m}\\ &\quad+\frac{1}{m}\sum_{k=1}^{m}\left|\frac{(y_k-a_k^\intercal\hat{\mu}_{c^*,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{c^*,k-1}a_k+\sigma^2}-\frac{(y_k-a_k^\intercal\mu_{c^*,k-1})^2}{a_k^\intercal\Sigma_{c^*,k-1}a_k+\sigma^2}\right|.\end{aligned}$$

Now we study bound for each individual term inside the sum over k. To simplify notation, we omit the dependence on k, c^* and $k-1$ without causing confusion. In the following let $z\triangleq y-a^\intercal\mu=a^\intercal(x-\mu)+w$, and let $\varrho\triangleq a^\intercal(\hat{\mu}-\mu)$. Hence, $|\varrho|\le|\beta|\cdot|\hat{\mu}-\mu|$ is bounded. Note that

$$\begin{aligned}&\left|\frac{(y-a^\intercal\hat{\mu})^2}{a^\intercal\widehat{\Sigma}a+\sigma^2}-\frac{(y-a^\intercal\mu)^2}{a^\intercal\Sigma a+\sigma^2}\right|\\ &\le\left|\frac{(y-a^\intercal\hat{\mu})^2}{a^\intercal\widehat{\Sigma}a+\sigma^2}-\frac{(y-a^\intercal\mu)^2}{a^\intercal\widehat{\Sigma}a+\sigma^2}\right|+\left|\frac{(y-a^\intercal\mu)^2}{a^\intercal\widehat{\Sigma}a+\sigma^2}-\frac{(y-a^\intercal\mu)^2}{a^\intercal\Sigma a+\sigma^2}\right|\\ &\le\frac{2|\varrho|\cdot|a^\intercal(x-\mu)+w-\varrho/2|}{\sigma^2}+(y-a^\intercal\mu)^2|\frac{\beta\delta}{\sigma^4}|\\ &=\frac{2|\varrho|\cdot|z-\varrho/2|}{\sigma^2}+|z|^2\frac{\beta\delta}{\sigma^4}\\ &\le\frac{|\varrho|^2+2|\varrho||z|}{\sigma^2}+|z|^2\frac{\beta\delta}{\sigma^4}.\end{aligned}\tag{26}$$

Since $m\le n$, from (26) we have

$$\begin{aligned}&\max_{k=1}^{m}\left|\frac{(y_k-a_k^\intercal\hat{\mu}_{c^*,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{c^*,k-1}a_k+\sigma^2}-\frac{(y_k-a_k^\intercal\mu_{c^*,k-1})^2}{a_k^\intercal\Sigma_{c^*,k-1}a_k+\sigma^2}\right|\\ &\le\max_{k=1}^{n}\left|\frac{(y_k-a_k^\intercal\hat{\mu}_{c^*,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{c^*,k-1}a_k+\sigma^2}-\frac{(y_k-a_k^\intercal\mu_{c^*,k-1})^2}{a_k^\intercal\Sigma_{c^*,k-1}a_k+\sigma^2}\right|\\ &\le\frac{1}{\sigma^4}\cdot\max_{k=1}^{n}\left\{\sigma^2(|\varrho_k|+2|z_k|)|\varrho_k|+\beta_k|z_k|^2\delta_{k-1}\right\}.\end{aligned}$$

Note that $z_k\triangleq a_k^\intercal(x-\mu_{c^*,k-1})+w_k\sim\mathcal{N}(a_k^\intercal(\mu_{c^*}-\mu_{c^*,k-1}),a_k^\intercal\Sigma_{c^*,k-1}a_k+\sigma^2)$, so for some $t\in(0,1)$, we have

$$\begin{aligned}&|z_k|\\ &<\frac{1}{t}\sqrt{[(\lambda_{c^*,k}+\delta_{k-1})\beta_k+\sigma^2]+(a_k^\intercal(\mu_{c^*}-\mu_{c^*,k-1}))^2}\triangleq\frac{b_k}{t}\end{aligned}$$

with probability exceeding $1-t^2$, where b_k is bounded.

Finally,

$$\frac{1}{m}\left|\sum_{k=1}^{m}(\sum_{\ell=1}^{C}\frac{(y_k-a_k^\intercal\hat{\mu}_{\ell,k-1})^2}{a_k^\intercal\widehat{\Sigma}_{\ell,k-1}a_k+\sigma^2}\cdot\hat{\pi}_\ell^k-\frac{(y_k-a_k^\intercal\mu_{c^*,k-1})^2}{a_i^\intercal\Sigma_{c^*,k-1}a_k+\sigma^2})\right|$$
$$\leq\hat{\eta}+\frac{2\ln|C|}{m}+\underbrace{\frac{1}{\sigma^4}\cdot\max_{k=1}^{n}\left\{\sigma^2(|\varrho_k|+2\frac{b_k}{t})|\varrho_k|+\beta_k\frac{b_k^2}{t^2}\delta_{k-1}\right\}}_{\eta_0}$$

with probability at least $1-nt^2$. Let

$$\Delta=\max_{k=1}^{n}\left\{\max(\varrho_k,\delta_{k-1})\right\},$$

and let

$$U=\frac{1}{\sigma^4}\cdot\max_{k=1}^{n}\left\{\sigma^2(\Delta+2nb_k)+\beta_kn^2b_k^2\right\}.$$

We choose $t=1/n$. Then

$$\eta_0=U\Delta.$$

Note that when there is no mismatch, $\delta_{k-1}=\varrho_k=0$ for $k\in[n]$, which leads to $\Delta=0$ and thus $\eta_0=0$. Here $\hat{\eta}=1/2$ is a parameter used in the multiplicative weight update algorithm. In particular, we can identify the correct component c^* with probability $1-1/n$ whenever $m=\mathcal{O}\left(\ln|C|/(\hat{\eta}+\eta_0)\right)$. For $k=1,\ldots,m$, we choose $\beta_k=1$. Thus, we need at most

$$\sum_{c=1}^{C}m_c+\mathcal{O}\left(\frac{\ln|C|}{\hat{\eta}+\eta_0}\right)$$

amount of power in total. □

Note that $|\varrho_k|$ can be computed recursively. We may derive a recursion. Let $z_k\triangleq a_k^\intercal(x-\mu_{k-1})+w_k=y_k-a_k^\intercal\mu_{k-1}$. Also Let $\varrho_k\triangleq a^\intercal(\hat{\mu}_k-\mu_k)$. Note that $\varrho_k=a^\intercal\xi_k$ for $\xi_k=\hat{\mu}_k-\mu_k$ in (17). Based on the recursion for ξ_k in (17) that we derived earlier, we have

$$\varrho_k=\frac{\sigma^2}{\beta_k\hat{\lambda}_k+\sigma^2}[\varrho_{k-1}+\frac{a_k^\intercal E_{k-1}a_k(y_k-a_k^\intercal\mu_{k-1})}{\beta_k\hat{\lambda}_k+\sigma^2-a_k^\intercal E_{k-1}a_k}]$$

and

$$|\varrho_k|\leq\frac{1}{\hat{\lambda}_k(\beta_k/\sigma^2)+1}[|\varrho_{k-1}|+\frac{\delta_k}{(\hat{\lambda}_k-\delta_k)+\sigma^2/\beta_k}|z_k|].$$

Proof of Lemma 3. The recursion of the diagonal entries can be written as

$$\Sigma_{ii}^{(k)}=\Sigma_{ii}^{(k-1)}-\frac{(\Sigma_{ij_{k-1}}^{(k-1)})^2}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}+\sigma^2/\beta_k}$$
$$=\frac{\Sigma_{ii}^{(k-1)}\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}(1-\rho_{ij_{k-1}}^{(k-1)})+\Sigma_{ii}^{(k-1)}\sigma^2/\beta_k}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}+\sigma^2/\beta_k}.$$

Note that for $i=j_{k-1}$,

$$\Sigma_{j_{k-1}j_{k-1}}^{(k)}=\frac{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}\sigma^2/\beta_k}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}+\sigma^2/\beta_k}\leq\frac{\gamma}{1+\gamma}\Sigma_{j_{k-1}j_{k-1}}^{(k-1)},$$

and for $i\neq j_{k-1}$,

$$\Sigma_{ii}^{(k)}\leq\frac{\Sigma_{ii}^{(k-1)}\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}(1-\rho^{(k-1)})+\Sigma_{ii}^{(k-1)}\sigma^2/\beta_k}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}+\sigma^2/\beta_k}$$
$$\leq\Sigma_{ii}^{(k-1)}\frac{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}(1-\rho^{(k-1)})+\sigma^2/\beta_k}{\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}+\sigma^2/\beta_k}$$
$$\leq\Sigma_{ii}^{(k-1)}\frac{1-\rho^{(k-1)}+\gamma}{1+\gamma}.$$

Therefore,

$$\mathrm{tr}(\Sigma_k)\leq(1-\frac{\rho^{(k-1)}}{1+\gamma})\mathrm{tr}(\Sigma_{k-1})-\frac{1-\rho^{(k-1)}}{1+\gamma}\Sigma_{j_{k-1}j_{k-1}}^{(k-1)}$$
$$\leq[1-\frac{(n-1)\rho^{(k-1)}+1}{n(1+\gamma)}]\mathrm{tr}(\Sigma_{k-1}).$$

□

Proof of Theorem 5. Let $\varepsilon\geq\sqrt{\|\Sigma_K\|\cdot\chi_n^2(p)}$, i.e. $\|\Sigma_K\|\leq\chi_{n,p,\varepsilon}$. Then Theorem 5 follows from

$$\begin{aligned}&\mathbb{P}_{x\sim\mathcal{N}(\mu_K,\Sigma_K)}[\|x-\mu_K\|_2\leq\varepsilon]\\&\geq\mathbb{P}_{x\sim\mathcal{N}(\mu_K,\Sigma_K)}[\|x-\mu_K\|_2\leq\sqrt{\|\Sigma_K\|\cdot\varepsilon^2}]\\&\geq\mathbb{P}_{x\sim\mathcal{N}(\mu_K,\Sigma_K)}[(x-\mu_K)^\intercal{\Sigma_K}^{-1}(x-\mu_K)\leq\chi_n^2(p)]=p.\end{aligned}\tag{27}$$

From Lemma 3, we have that when the powers β_i are sufficiently large

$$\|\Sigma_K\|\leq\mathrm{tr}(\Sigma_K)\leq(1-\frac{1}{n(1+\gamma)})^K\mathrm{tr}(\Sigma).$$

Hence for (27) to hold, we can simple require $(1-\frac{1}{n(1+\gamma)})^K\mathrm{tr}(\Sigma)\leq\chi_{n,p,\varepsilon}$, or equivalently (15) in Theorem 5. □